Die Fehlalarmfalle
Warum hat ein guter Test so miserable Ergebnisse?

Klaus Pommerening

Mai 2020

Das Problem

Nehmen wir an, jemand hat einen Labortest für eine Infektion, nennen wir sie CoV-2, entwickelt. Dieser Test hat hervorragende Qualitätskennzahlen:

- Er erkennt bei 98 % der Infizierten die Infektion korrekt. Der Fachausdruck dafür ist **Sensitivität** – diese ist also[1] 98 %.
- Er schlägt nur bei 5 % von Nichtinfizierten an, erkennt also die Nichtinfizierten zu 95 % richtig. Der Fachausdruck dafür ist **Spezifität** – diese ist also 95 %.

Nun wird eine Person getestet, und der Test fällt positiv aus. *Wie groß ist die Wahrscheinlichkeit, dass diese Person infiziert ist?* 95 % oder 98 %?

Weit gefehlt – keine der beiden Antworten kommt auch nur in die Nähe der richtigen Lösung!

Ein Zahlenbeispiel

In Wirklichkeit kann die Frage ohne eine weitere Information gar nicht beantwortet werden, eine Information, die mit der Qualität des Tests *nichts zu tun hat*, für seine Ergebnisse trotzdem entscheidend ist: Wieviele Personen, egal ob schon getestet oder nicht, sind insgesamt infiziert?

Nehmen wir an, von der Gesamtbevölkerung von 80 Millionen eines Staates seien 160 000 aktuell infiziert. Der Fachausdruck dafür ist **Prävalenz** – diese ist in diesem Zahlenbeispiel also 160 000 geteilt durch 80 Millionen, macht 0,002, das sind 2 ‰.

Welche Ergebnisse können wir dann bei einem Test von 100 000 Personen erwarten? Von diesen sind wahrscheinlich etwa 200 infiziert. Bei diesen ist der Test wahrscheinlich in 196 Fällen positiv, in vier Fällen negativ.

Die übrigen 99 800 unter den Testpersonen sind von der Infektion verschont. Diese bekommen zu 95 % das korrekte negative Testergebnis, bei den übrigen 5 % ist das Ergebnis positiv, also bei 4 990, gegenüber der überwältigenden Mehrheit von 94 810 korrekt negativen Ergebnissen.

[1] Man ist natürlich versucht zu fragen: Warum nicht 100 %? Nun, erstens können Fehler auftreten, angefangen von der Probenentnahme bis hin zur Fehleinschätzung einer Färbung im Reagenzglas. Außerdem nimmt man bei höherer Sensitivität in der Regel in Kauf, dass mehr Nichtinfizierte fälschlicherweise ein positives Testergebnis erhalten, dass also die Spezifität schlechter ist. Ideal wäre natürlich ein Test, der Infizierte und Nichtinfizierte sicher auseinanderhalten kann, also jeweils 100 % Sensitivität und Spezifität hat.

Um die Auswirkung dieser Überlegungen deutlich zu erkennen, werden die in etwa zu erwartenden Ergebnisse in eine **Vierfeldertafel** eingetragen:

CoV2-Test	infiziert? ja	nein	Summe
+	196	4 990	5 186
–	4	94 810	94 814
Summe	200	99 800	100 000

Wir haben also 5 186 positive Testergebnisse zu erwarten, darunter aber nur 196 korrekte bei 4 990 (!) falschen. Das bedeutet (in diesem Zahlenbeispiel):

> *Die Wahrscheinlichkeit, dass eine positiv getestete Person tatsächlich infiziert ist, ist 196/5 186, also ungefähr 3,8%.*

Diese Größe nennt man den **positiven Vorhersagewert.**

> *Die ungefähr 96,2% positiven Ergebnisse des Tests bei Nichtinfizierten bedeuten hingegen jeweils einen* **Fehlalarm.**

Erklärung

Woran liegt dieses „Versagen“? Nun, die vielen, nämlich fast 5 000, „falsch positiven“ Testergebnisse kamen einfach dadurch zustande, dass die Anzahl der Nichtinfizierten so viel größer ist als die Zahl der Infizierten und damit der geringe Prozentsatz an falsch positiven Ergebnissen trotzdem in absoluten Zahlen beträchtlich ist – jedenfalls viel größer als die Zahl der mit hoher Wahrscheinlichkeit gefundenen „echt positiven“ Ergebnisse.

Der Schluss von der Ursache auf die Wirkung lässt sich eben nicht einfach umkehren zu einem Schluss von der Wirkung auf die Ursache! Mehr dazu – mit etwas mathematischer Untermauerung – im letzten Abschnitt.

Sehen wir uns als Gegenprobe ein Zahlenbeispiel an, bei dem ungefähr die Hälfte der untersuchten Bevölkerung infiziert ist. Dann sähe die Vierfeldertafel (bei gleicher Sensitivität und Spezifität des Tests) etwa so aus:

Test	infiziert? ja	nein	Summe
+	49 000	2 500	51 500
–	1 000	47 500	48 500
Summe	50 000	50 000	100 000

Von den 51 500 postiv getesteten wären also 49 000 tatsächlich infiziert – der positive Vorhersagewert läge bei 49 000/51 500, also bei 95 %. Die Fehlalarmquote in diesem Beispiel wäre 2 500/51 500, also etwa 5 %. Das sähe akzeptabel aus.

Der Haken bei der Sache ist nur, dass (glücklicherweise) eine so hohe Zahl von tatsächlich Infizierten in der Praxis so gut wie nie vorkommt.

Wichtig zu wissen ist jedenfalls, dass der miserable Vorhersagewert nicht ein Qualitätsmanko des Tests ist, sondern *durch die geringe Prävalenz der nachzuweisenden Infektion bedingt ist*. Unter einer solchen Fehleinschätzung leidet immer wieder das Ansehen vieler wichtiger Tests und Diagnoseverfahren, z. B. des Mammographie-Screenings.

Auswege aus der Falle?

Einen einfachen Ausweg gibt es nicht, die Zahlen sprechen für sich. Je seltener eine Infektion oder Erkrankung in der Bevölkerung ist, desto mehr Fehlalarme, also falsch positive Testergebnisse, muss man in Kauf nehmen. Man könnte eventuell diese Zahl senken, indem man die Spezifität des Tests erhöht. Man muss dann aber in Kauf nehmen, dass er mehr falsch negative Ergebnisse produziert, dass einem also Infizierte „durch die Lappen“ gehen. Das dürfte in den meisten Fällen keine reale Option sein, denn das hauptsächliche Ziel ist ja gerade, die Infizierten zu finden.

Nehmen wir trotzdem an, die Spezifität könnte erhöht werden. Das könnte etwa geschehen, indem man einen Grenzwert in Richtung „weniger positive Ergebnisse“ verschiebt – z. B., wenn man den Grad der Färbung in einem Reagenzglas, der einen positiven Befund anzeigen soll, etwas höher ansetzt. Könnte man bei unserem Zahlenbeispiel eine Spezifität von 98 % bei einer Sensibilität von 95 % erreichen, so sähe die Ergebnistafel so aus:

CoV2-Test	infiziert? ja	nein	Summe
+	190	1 996	2 186
–	10	97 804	97 814
Summe	200	99 800	100 000

Die Fehlalarmrate dieses modifizierten Tests wäre dann 1 996/2 186, also immer noch über 91 %, und der Vorhersagewert wäre knapp 9 %. Der Preis dafür wäre allerdings, dass 10 statt nur 4 wirklich Infizierte übersehen werden.

Was bleibt an möglichen Auswegen?

- Natürlich hilft es, wenn ein besserer Test gefunden werden kann, also mit erhöhter Spezifität, und dabei mindestens gleich guter Sensitivität. Das ist aber höchstens dann eine realistische Option, wenn es einen solchen besseren Test schon gibt, er aber wegen seiner Verfügbarkeit, seiner Belastung für die zu testenden Personen oder seiner Kosten zunächst nicht eingesetzt wird. Und wie gesehen ist der Vorhersagewert trotzdem nicht wirklich viel besser, weil er auch für den besseren Test sehr stark von der „Durchseuchung“ der Bevölkerung abhängt.

- Das wichtigste ist auf jeden Fall, dass die Ärztin, die dem Betroffenen das Testergebnis mitteilt, sich der Problematik bewusst ist und den Vorhersagewert in etwa kennt. Sie sollte dem Patienten also nicht sagen: „Sie haben Krebs.“ – sondern: „Das Testergebnis ist positiv, aber die Wahrscheinlichkeit, dass Sie wirklich Krebs haben, liegt bei etwa 4 %. Wir sollten das natürlich weiter überprüfen.“

- Deutlich höher ist der Vorhersagewert, wenn aus anderem Zusammenhang schon ein Verdacht auf die Infektion besteht, wenn der Test also nur bei einem Teil der Bevölkerung

durchgeführt wird, bei dem die Infektion wesentlich häufiger auftritt. Bei einer Infektion, die durch die Atemluft auf Personen in geringem Abstand übertragen wird, würden beispielsweise nur Fälle getestet, die direkten Kontakt mit einem bereits Infizierten hatten.

Füllen wir für diese Situation einer „vorgesiebten“ Testpopulation wieder eine Vierfeldertafel aus. Dabei wird angenommen, dass das etwa 10 000 Personen sind und unter diesen etwa 200, also 2 % statt nur 2 ‰infiziert sind[2].

	infiziert?		
Test	ja	nein	Summe
+	196	490	686
–	4	9 310	9 314
Summe	200	9 800	10 000

Von den 686 positiv getesteten sind 196 wirklich infiziert, d. h., der Vorhersagewert ist jetzt 196/686, liegt also bei knapp 29 %. *Das ist immer noch sehr bescheiden, aber doch deutlich besser als der ursprüngliche Wert von 3,8 %.*

An der Zahl der 4 nicht entdeckten Infizierten hat sich übrigens nichts geändert – diese kommt ja durch die Sensitivität von 98 % zustande.

Die Lehre aus dieser Analyse mit Zahlenbeispiel ist also:

> *Ein positives Testergebnis ist meistens noch weit von einer endgültigen Diagnose entfernt. Es muss so gut wie immer durch weitere Untersuchungen überprüft werden.*

Was bringt ein zweiter Test?

Das Szenario „Erst ein Vortest, dann ein weiterer Test“ passt auch für den Fall, dass es zwei unabhängige Tests für die Infektion gibt. Dann kann das Ergebnis von Test 1 als Vorauswahl für Test 2 dienen. Die positiv Getesteten von Test 1 bilden in diesem Szenarion die zu untersuchende Bevölkerung von Test 2, und in dieser Gruppe ist die Prävalenz wesentlich höher – im obigen Zahlenbeispiel 196 von 5 186, also[3] 3,8 %. Das ist entscheidend mehr als die Prävalenz von 2 ‰ in der Gesamtbevölkerung.

Füllen wir für die Situation eines zweiten, unabhängigen Tests auch wieder eine Vierfeldertafel aus, für die 5 186 Testpersonen, die beim Test 1 auffällig waren. Ansonsten gehen wir für das Beispiel als Qualitätskenngrößen des zweiten Tests von 95 % Sensitivität und 98 % Spezifität aus, bei einer Prävalenz von 3,8 % in der untersuchten Gruppe:

	infiziert?		
Test	ja	nein	Summe
+	186	100	286
–	10	4890	4 900
Summe	196	4 990	5 186

[2]Die Vorauswahl hat also alle Infizierten erfasst. Das entspricht einem „Vortest“ mit 100 % Sensitivität, aber einer sehr niedrigen Spezifität von nur 2 %.

[3]Es ist kein Zufall, dass das genau mit dem Vorhersagewert von Test 1 übereinstimmt.

Von den 286 positiv getesteten sind 186 wirklich infiziert, d. h., der Vorhersagewert ist jetzt 186/286, liegt also bei 65 %. Das ist wesentlich besser, hat aber einen Preis: Weitere 10 Infizierte sind als nichtinfiziert klassifiziert worden. *Das ist in der Regel nicht hinnehmbar. Wenn ein zweiter Test oder eine weitere Untersuchung angeschlossen wird, sollte diese eine Sensitivität von 100 % haben!* Jede Abweichung davon lässt weitere Infizierte unentdeckt.

Im nächsten und letzten Abschnitt werden diese an Zahlenbeispielen gewonnenen Erkenntnisse mit etwas mehr Mathematik unterfüttert.

Etwas Theorie: Die Formel von Bayes

Das Problem des Umkehrschlusses von der Wirkung auf die Ursache wird in der Wahrscheinlichkeitsrechnung durch die Formel von Bayes beschrieben. Zum Verständnis ist Vertrautheit mit dem (aus der Schulmathematik bekannten) „algebraischen“ Ansatz, unbestimmte Zahlenwerte durch Buchstaben wiederzugeben, nötig. Außerdem wird der mathematische Begriff der Wahrscheinlichkeit benötigt, wobei hier die naive Vorstellung von Wahrscheinlichkeit als relativer Häufigkeit ausreicht[4].

Die Überlegung startet wieder mit der Vierfeldertafel der Testergebnisse, diesmal mit unbestimmten Werten ausgefüllt:

	infiziert?		
Test	ja	nein	Summe
+	a	b	$a+b$
−	c	d	$c+d$
Summe	$a+c$	$b+d$	n

Dabei ist $n = a+b+c+d$ die Gesamtzahl der untersuchten Population. Die relevanten Größen werden jetzt durch Wahrscheinlichkeiten ausgedrückt. Es steht I für das Ereignis „infiziert“, und T für das Ereignis „Testergebnis positiv“.

Prävalenz ist die Wahrscheinlichkeit $P(I)$ für I in der Population, hier ausgedrückt durch die relative Häufigkeit $(a+c)/n$.

[4]Wir stellen uns eine Menge aus n Elementen als Sammlung von Kugeln in einer Urne vor, rote und weiße. Die Teilmenge $B \subseteq A$ soll genau aus den m roten Kugeln bestehen. Dann ist die Wahrscheinlichkeit $P(B)$, bei zufälliger Auswahl einer Kugel eine rote zu erwischen, gerade die relative Häufigkeit m/n.

Sensitivität ist die bedingte Wahrscheinlichkeit $P(T|I)$ eines positiven Testergebnisses unter der Voraussetzung, dass die Infektion vorliegt, hier ausgedrückt durch die relative Häufigkeit $a/(a+c)$.

Spezifität ist die bedingte Wahrscheinlichkeit[5] $P(\neg T|\neg I)$ eines negativen Testergebnisses unter der Voraussetzung, dass keine Infektion vorliegt, hier ausgedrückt durch die relative Häufigkeit $d/(b+d)$.

Vorhersagewert ist die bedingte Wahrscheinlichkeit $P(I|T)$ dafür, dass bei positivem Testergebnis tatsächlich die Infektion vorliegt. Er wird hier ausgedrückt durch die relative Häufigkeit $a/(a+b)$.

Der Vorhersagewert ist also die Wahrscheinlichkeit für das Zutreffen des Umkehrschlusses von der Wirkung auf die Ursache. Die Formel von Bayes drückt ihn durch die übrigen Größen aus:

$$P(I|T) = \frac{P(T|I) \cdot P(I)}{P(T)} \tag{1}$$

Das ist kein tiefliegender mathematischer Satz, sondern folgt direkt aus der Definition der bedingten Wahrscheinlichkeit:

$$P(T|I) = \frac{P(T \cap I)}{P(I)} \quad \text{und} \quad P(I|T) = \frac{P(T \cap I)}{P(T)}\,,$$

wobei $P(T \cap I)$ die Wahrscheinlichkeit dafür ist, dass T und I gemeinsam auftreten.

Schöner wäre es, wenn auf der rechten Seite der Formel (1) nur die Größen Prävalenz, Sensitivität und Spezifität aufträten. Das kann man leicht erreichen, indem man die „störende“ Wahrscheinlichkeit $P(T)$ für ein positives Testergebnis durch diese Größen ausdrückt: Da T die disjukte Vereinigung $(T \cap I) \cup (T \cap \neg I)$ ist, ist

$$P(T) = P(T|I) \cdot P(I) + P(T|\neg I) \cdot P(\neg I)$$

und dabei $P(\neg I) = 1 - P(I)$ und $P(T|\neg I) = 1 - P(\neg T|\neg I)$. Setzt man dies in die Formel (1) ein, so erhält man eine alternative Formel, auf deren rechter Seite nur noch die erwünschten Größen auftreten[6]:

$$P(I|T) = \frac{P(T|I) \cdot P(I)}{P(T|I) \cdot P(I) + (1 - P(\neg T|\neg I)) \cdot (1 - P(I))}\,, \tag{2}$$

die sich allerdings nicht durch besondere Übersichtlichkeit auszeichnet. In Worten ausgedrückt:

$$\text{Vorhersagewert} = \frac{\text{Sensitivität} \cdot \text{Prävalenz}}{\text{Sensitivität} \cdot \text{Prävalenz} + (1 - \text{Spezifität}) \cdot (1 - \text{Prävalenz})}\,.$$

[5] Das Symbol ¬ bedeutet die logische Negation.

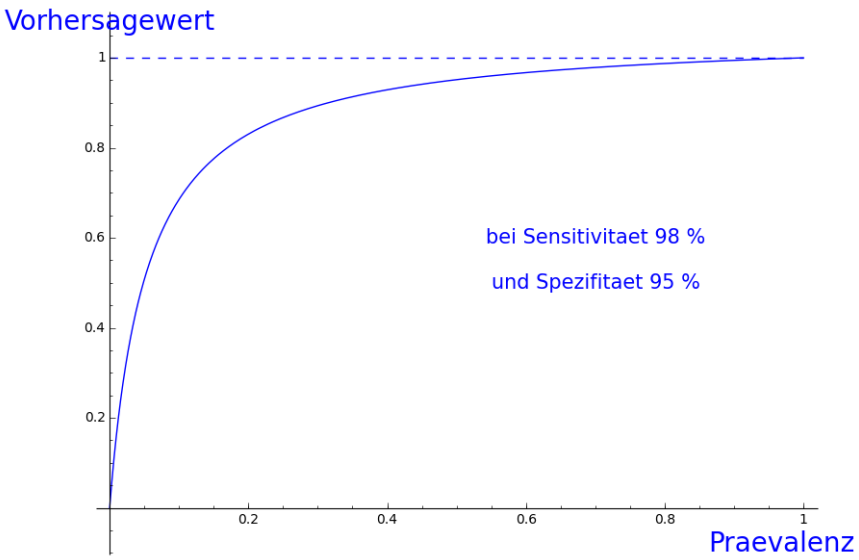
[6] Man kann die Richtigkeit der Formel auch anhand der relativen Häufigkeiten nachvollziehen: In der Formel steht dann $\frac{a}{a+c}$ für die Sensitivität $P(T|I)$ und $\frac{a+c}{n}$ für die Prävalenz $P(I)$, also $\frac{a}{n}$ für das Produkt $P(T|I) \cdot P(I)$, das sowohl im Zähler als auch im Nenner steht. Im Nenner kommt dann noch die Spezifität $P(\neg T|\neg I)$ vor, die durch $\frac{d}{b+d}$ repräsentiert wird. Damit wird $1 - P(\neg T|\neg I)$ zu $1 - \frac{d}{b+d} = \frac{b}{b+d}$, und weil $1 - P(I)$ zu $1 - \frac{a+c}{n} = \frac{b+d}{n}$ wird, wird das Produkt $(1 - P(\neg T|\neg I)) \cdot (1 - P(I))$ im Nenner zu $\frac{b}{b+d} \cdot \frac{b+d}{n} = \frac{b}{n}$. Der gesamte Nenner wird also zu $\frac{a}{n} + \frac{b}{n} = \frac{a+b}{n}$, der ganze Bruch also zu $\frac{a}{n} / \frac{a+b}{n} = \frac{a}{a+b}$, was ja gerade für den Vorhersagewert steht.

Wirklich übersichtlich ist das auch nicht. Mathematiker behelfen sich in solchen Situationen gerne, indem sie einbuchstabige Symbole verwenden. Also etwa v für den Vorhersagewert, q für die Prävalenz, r für die Sensitivität und s für die Spezifität. Dann bekommt die Formel von Bayes die Gestalt

$$v = \frac{r \cdot q}{r \cdot q + (1 - s) \cdot (1 - q)}\,, \tag{3}$$

an der man die Abhängigkeit von v von den anderen drei Größen q, r und s ablesen kann.

Als Beispiel betrachten wir v als Funktion von q bei festen r und s, also die funktionale Abhängigkeit des Vorhersagewerts von der Prävalenz für einen Test mit bekannter Sensitivität und Spezifität.



Die Abbildung zeigt diese Funktion für das Beispiel mit $r = 98\,\%$ und $s = 95\,\%$. Man erkennt, dass v als Funktion von q zwischen 0 und 1 monoton von 0 bis 1 zunimmt. Ab einer Prävalenz von ungefähr 0.2, also $20\,\%$, liegt der Vorhersagewert mindestens bei 0.8, also $80\,\%$, was man als einigermaßen akzeptabel ansehen würde. In der Praxis ist die Prävalenz glücklicherweise meistens sehr viel niedriger, was allerdings *den unerwünschten, aber unvermeidbaren Nebeneffekt eines miserabel niedrigen Vorhersagewerts mit sich bringt.*

Noch eine letzte Bemerkung: Wie sicher kann sich eine negativ Getestete sein, dass sie nicht infiziert ist? Dies wird durch den **negativen Vorhersagewert** $P(\neg I|\neg T)$ ausgedrückt. In der Vierfeldertafel entspricht das dem Verhältnis $d/(c + d)$ und hat im ersten Zahlenbeispiel den Wert 94 810/94 814, der erfreulicherweise praktisch $100\,\%$ bedeutet.